Quality-Driven Machine Learning-based Data Science Pipeline Realization: a software engineering approach

Giordano d'Aloisio
giordano.daloisio@graduate.univaq.it
University of L'Aquila, Italy

ABSTRACT

The recently wide adoption of data science approaches to decision making in several application domains (such as health, business and even education) open new challenges in engineering and implementation of this systems. Considering the big picture of data science, Machine learning is the wider used technique and due to its characteristics, we believe that a better engineering methodology and tools are needed to realize innovative data-driven systems able to satisfy the emerging quality attributes (such as, debias and fariness, explainability, privacy and ethics, sustainability). This research project will explore the following three pillars: *i)* identify key quality attributes, formalize them in the context of data science pipelines and study their relationships; *ii)* define a new software engineering approach for data-science systems development that assures compliance with quality requirements; *iii)* implement tools that guide IT professionals and researchers in the realization of ML-based data science pipelines since the requirement engineering. Moreover, in this paper we also presents some details of the project showing how the feature models and model-driven engineering can be leveraged to realize our project.

CCS CONCEPTS

• **Software and its engineering → Extra-functional properties**; *Designing software*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

machine learning, pipelines, software quality, product-line architecture, model-driven

ACM Reference Format:
Giordano d'Aloisio. 2022. Quality-Driven Machine Learning-based Data Science Pipeline Realization: a software engineering approach. In *44th International Conference on Software Engineering Companion (ICSE '22 Companion), May 21–29, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3510454.3517067

1 INTRODUCTION

Data Science (DS), and in particular Machine Learning (ML), systems are increasingly becoming a used instrument, applied to all application domains and affecting our real life. Such systems can be defined as a set of one or more pipelines (or workflows), which take as input raw (unprocessed) data and returns actionable answers to questions in the form of machine learning models. In this paper, we focus on DS pipelines that leverage on ML, that we call ML pipelines.

Problem definition. A generic ML pipeline is depicted in Figure 1 [1, 18], where rounded blue boxes represent phases and yellow squared boxes represent quality attributes that influence the quality of the final system. The quality attributes are reported in correspondence of the tasks that affect them. As an example, if the system requires high fairness (e.g. for legal reasons [4, 7, 14]), then debiasing component must be included during the extraction and cleaning or the analysis phase in order to achieve fairness [16]. Some quality attributes can affect other qualities, for example, fairness usually has a negative influence on the predictions' performance [8, 12]. So, if the system is required to have also a high prediction performance, other solutions must be considered during this step. Hence, the identification and formalization of quality attributes and requirements, and their handling in the development of ML pipelines are complex tasks, also given the complex influence among them.

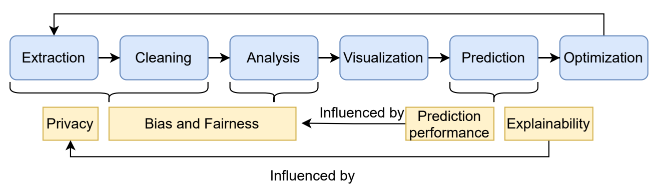

Figure 1: Data Science workflow with involved quality attributes

Expected contribution. In this research project, we aim to formalize a common definition of quality in ML pipelines taking into account the influence among different quality attributes. Moreover, leveraging on such definition we aim to realize a model-driven framework that automatically generates ML pipelines compliant with quality requirements. In particular, the expected contribution is given by answering the following research questions:

> **RQ 1)** How can we formalize quality attributes and requirements of the whole ML pipeline?
> **RQ 2)** How can we formalize how quality attributes influence each other?
> **RQ 3)** How can data scientists specify quality requirements with a user-friendly formalism?
> **RQ 4)** How can we automatically generate the ML pipeline assuring the functional and quality requirements?

ICSE '22 Companion, May 21–29, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9223-5/22/05…$15.00
https://doi.org/10.1145/3510454.3517067

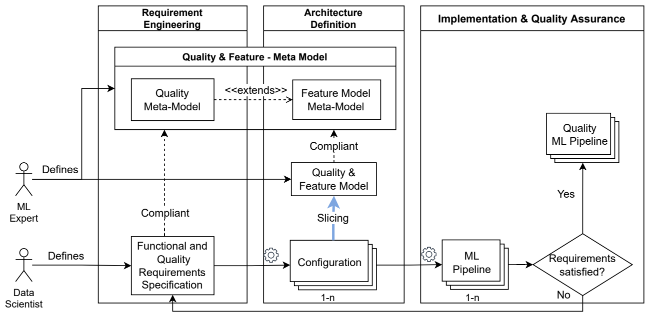

Figure 2: High-level architecture of the proposed framework

2 RELATED WORK

The quality assurance (QA) problem in ML pipelines has acquired much relevance in the last years. Many articles highlight the needing to define and formalize new standard quality requirements for such systems [5, 6, 9, 20]. Several solutions have been proposed to formalize and identify standard quality requirements in ML pipelines. Concerning the standardization of workflows and quality requirements, Studer et al. proposed the *CRISP_ML* process model to develop quality machine learning systems [18]. Instead, the *Q4AI* consortium proposed a set of guidelines for the quality assurance of ML systems for specific domains [10]. Concerning the modelling of quality requirements, Azimi et al. proposed a layered model for the QA of Machine Learning systems in the context of IoT [3], instead Ishikawa proposed a framework for the quality evaluation of an ML system using an argumentation approach [11]. Finally, Siebert et al. presented a meta-model for the formal definition of quality requirements in ML pipelines [17]. To the best of our knowledge, this is the first attempt at formalizing the quality requirements of ML pipelines using a model-driven approach.

From this analysis of related work, we can conclude that there is a robust research motivation in formalizing and defining new quality requirements for ML systems. Many attempts have been proposed to solve these issues, and several definitions of quality requirements, metrics and components are now available from the literature [17, 18, 20]. However, these issues are still not fully addressed: *i)* a general and formal definition of quality in ML pipelines is still missing *ii)* the analyzed papers do not cover the entire development process of ML pipelines starting from requirement specification to system implementation and quality assurance *iii)* a formalization of the influence among quality attributes is still missing. With this research project, we aim to overcome these issues.

3 PROPOSED APPROACH

The proposed approach aims to define an innovative model-driven framework that guides data scientists in the development of ML pipelines assuring quality requirements. Figure 2 depicts the high-level architecture of such a framework. As discussed earlier (see Figure 1), ML pipelines are made of common phases [1, 18] that embed a set of standard components identified by the system's functional requirements (like the ML model suited for a ML goal such as classification), and a set of variability points that represents different methods to implement the functional requirements and to satisfy quality constraints. For example, if we consider classification problem and explainability quality, variability points are represented by models as *Support Vector Machines* combined with explainability methods like *Local Interpretable Model-Agnostic Explanations (LIME)* [15], or *Decision Trees*. While Support Vector Machines needs to be combined with a specific explainability method, Decision Tree must not because it has an intrinsic explainability that guarantees the quality requirement.

Product-Line Architectures, specified by Feature Models, [13, 19] represent a suitable model to formalize ML pipelines with variability. But, they miss adequate means to specify quality attributes and requirements. In fact, they do not allow to specify thresholds and metrics. To address this issue, in our approach (see Figure 2), we propose to extend the feature models meta-model to enable: *i)* the creation, by the machine learning expert, of an enriched feature model with associated quality attributes (as done in [2]) *ii)* the specification of functional and quality requirements by the data scientist. In particular, during the *Requirement Engineering*, the data scientist specifies a set of functional and quality requirements compliant to the defined meta-model. These requirements are used during *Architecture Definition* to automatically generate, from the extended feature model provided by the machine learning expert, a set of ML pipeline configurations able to satisfy the defined functional requirements (the *Configuration* boxes in the figure). The configurations are defined by removing from the feature model all the components (and their relative specification) not suitable to meet the specified requirements.

The generated configurations are given as input to the *Implementation Quality Assurance* step that aims to : *i)* generate for each configuration a python script implementing the *ML Pipeline*, and *ii)* verify that at least one generated pipeline satisfies the quality requirements (namely, *Quality ML Pipeline*).

The framework returns the set of Quality ML Pipelines, satisfying quality constraints, if any, ordemands the data scientist to relax quality requirements and repeat the process.

4 EVALUATION PLAN

The research project is in its infancy and we plan to answer **RQ 1** and **RQ 2** during the end of the first year of the PhD. In particular, we plan to better study the quality attributes of ML-based DA systems and their influences and to define how to embed quality requirements and characteristics in a product line architecture. We plan to present these results to the community through an extended version of feature meta-model. The second and third year will be focused on answering **RQ 3** and **RQ 4** with the implementation of a model-driven framework for the automatic implementation of ML pipelines assuring quality requirements. In order to ease the requirement specification task, we also plan to define a Domain Specific Language. We aim to test the final application by creating ML systems following quality requirements like fairness, explainability, privacy of sensitive information. In order to prove the generality of the approach and test different quality attributes, we will consider use cases from several application domains.

Acknowledgments. This work is partially supported by Territori Aperti a project funded by Fondo Territori Lavoro e Conoscenza CGIL CISL UIL and by SoBigData-PlusPlus H2020-INFRAIA-2019-1 EU project, contract number 871042.

REFERENCES

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[2] Mohsen Asadi, Samaneh Soltani, Dragan Gasevic, Marek Hatala, and Ebrahim Bagheri. 2014. Toward automated feature model configuration with optimizing non-functional requirements. *Information and Software Technology* 56, 9 (Sept. 2014), 1144–1165. https://doi.org/10.1016/j.infsof.2014.03.005

[3] Shelernaz Azimi and Claus Pahl. 2020. A Layered Quality Framework for Machine Learning-driven Data and Information Models.. In *ICEIS (1)*. 579–587.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *https://doi.org/10.1177/0049124118782533* 50, 1 (2018), 3–44. https://doi.org/10.1177/0049124118782533 Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[5] Gurol Canbek. 2021. The need for a systematic machine-learning process: A proposal via a mobile malware classification case study. In *2021 International Conference on Information Security and Cryptology (ISCTURKEY)*. IEEE, Ankara, Turkey, 173–178. https://doi.org/10.1109/ISCTURKEY53027.2021.9654378

[6] Elizamary de Souza Nascimento, Iftekhar Ahmed, Edson Oliveira, Márcio Piedade Palheta, Igor Steinmacher, and Tayana Conte. 2019. Understanding development process of machine learning systems: Challenges and solutions. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 1–6.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (2012), 214–226. https://doi.org/10.1145/2090236.2090255

[8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Sydney NSW Australia, 259–268. https://doi.org/10.1145/2783258.2783311

[9] Görkem Giray. 2021. A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software* (2021), 111031.

[10] Koichi Hamada, Fuyuki Ishikawa, Satoshi Masuda, Tomoyuki Myojin, Yasuharu Nishi, Hideto Ogawa, Takahiro Toku, Susumu Tokumoto, Kazunori Tsuchiya, Yasuhiro Ujita, et al. 2020. Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence.. In *SEKE*. 335–341.

[11] Fuyuki Ishikawa. 2018. Concepts in quality assessment for machine learning-from test data to arguments. In *International Conference on Conceptual Modeling*. Springer, 536–544.

[12] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[13] Kyo C Kang, Sholom G Cohen, James A Hess, William E Novak, and A Spencer Peterson. 1990. *Feature-oriented domain analysis (FODA) feasibility study*. Technical Report. Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst.

[14] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems* (2017), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[15] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (Dec. 2020), 18. https://doi.org/10.3390/e23010018

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 1–35. https://doi.org/10.1145/3457607

[17] Julien Siebert, Lisa Joeckel, Jens Heidrich, Adam Trendowicz, Koji Nakamichi, Kyoko Ohashi, Isao Namba, Rieko Yamamoto, and Mikio Aoyama. 2021. Construction of a quality model for machine learning systems. *Software Quality Journal* (2021), 1–29.

[18] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. 2021. Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction* 3, 2 (2021), 392–413.

[19] Thomas Thum, Christian Kastner, Sebastian Erdweg, and Norbert Siegmund. 2011. Abstract features in feature modeling. In *2011 15th International Software Product Line Conference*. IEEE, 191–200.

[20] Hugo Villamizar, Tatiana Escovedo, and Marcos Kalinowski. 2021. Requirements Engineering for Machine Learning: A Systematic Mapping Study. In *SEAA*. 29–36.